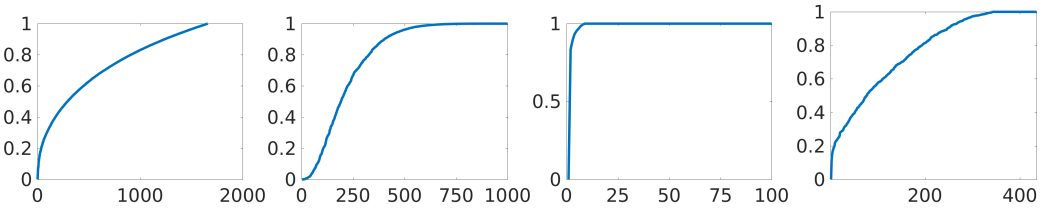
SUPPLEMENTARY MATERIALS

1. Spectral Properties of L. The spectral properties of L are relevant to the spectral projection and approximation algorithms from the previous section. Figure 1 shows the spectra for our four examples. Note that in all cases the spectrum is contained in the interval $[0, 2]$, consistent with the theoretical result in [1, Lemma 1.7, Chapter 1]. The size of the eigenvalues near to 0 will determine the accuracy of the spectral projection algorithm. The rate at which the spectrum accumulates at a value near 1, an accumulation which happens for all but the MNIST data set in our four examples, affects the accuracy of the spectral approximation algorithm. There is theory that goes some way towards justifying the observed accumulation; see [2, Proposition 9, item 4]. This theory works under the assumption that the features x_j are i.i.d samples from some fixed distribution, and the graph Laplacian is constructed from weights $w_{ij} = k(x_i, x_j)$, and k satisfies symmetry, continuity and uniform positivity. As a consequence the theory does not apply to the graph construction used for the MNIST dataset since the K-nearest neighbor graph is local; empirically we find that this results in a graph violating the positivity assumption on the weights. This explains why the MNIST example does not have a spectrum which accumulates at a value near 1. In the case where the spectrum does accumulate at a value near 1, the rate can be controlled by adjusting the parameter τ appearing in the weight calculations; in the limit $\tau = \infty$ the graph becomes an unweighted complete graph and its spectrum comprises the the two points $\{0, \lambda\}$ where $\lambda \to 1$ as $n \to \infty$ – see Lemma 1.7 in Chapter 1 of [1].



(a) MNIST49 (b) Two Moons (c) Hyperspectral (d) Voting Records

FIG. 1. *Spectra of graph Laplacian of various datasets. See main text for the description of the datsets and graph construction parameters.*

2. MAP Estimation as Semi-supervised Classification Method. We first prove the convexity of the probit negative log likelihood.

PROPOSITION 1. *Let* $\mathsf{J}_{\mathrm{p}}(u)$ *be the MAP estimation function for the probit model:*

$$\mathsf{J}_{\mathrm{p}}(u) = \frac{1}{2}\langle u, Pu\rangle - \sum_{j\in Z'} \log\Big(\Psi(y(j)u(j);\gamma)\Big).$$

If $y(j) \in \{\pm 1\}$ *for all* j *then* J_{p} *is a convex function in the variable* u.

Proof. Since P is semi positive definite, it suffices to show that

$$\sum_{j\in Z'} \log\Big(\Psi(y(j)u(j);\gamma)\Big)$$

is convex. Thus, since $y(j) \in \{\pm 1\}$ for all j, it suffices to show that $\log\Big(\Psi(x;\gamma)\Big)$ is concave with respect to x. Since

$$\Psi(x;\gamma) = \frac{1}{\sqrt{2\pi\gamma}} \int_{-\infty}^{x} \exp(\frac{-t^2}{2\gamma^2})dt,$$

we have $\Psi(\gamma x;\gamma) = \Psi(x;1)$. Since scaling x by a constant doesn't change convexity, it suffices to consider the case $\gamma = 1$. Taking the second derivative with $\gamma = 1$, we see that it suffices to prove that, for all $x \in \mathbb{R}$ and all $\gamma > 0$,

$$\Psi^{''}(x;1)\Psi(x;1) - \Psi^{'}(x;1)\Psi^{'}(x;1) < 0. \tag{1}$$

Plugging in the definition of Ψ, we have

$$\Psi^{''}(x;1)\Psi(x;1) - \Psi^{'}(x;1)\Psi^{'}(x;1) = \frac{-1}{2\pi}\exp(\frac{-x^2}{2})\Big(x\int_{-\infty}^{x}\exp(\frac{-t^2}{2})dt + \exp(\frac{-x^2}{2})\Big). \tag{2}$$

Clearly the expression in equation (2) is less than 0 for $x \geq 0$. For the case $x < 0$, divide equation (2) by $\frac{1}{2\pi}\exp(\frac{-x^2}{2})$ and note that this gives

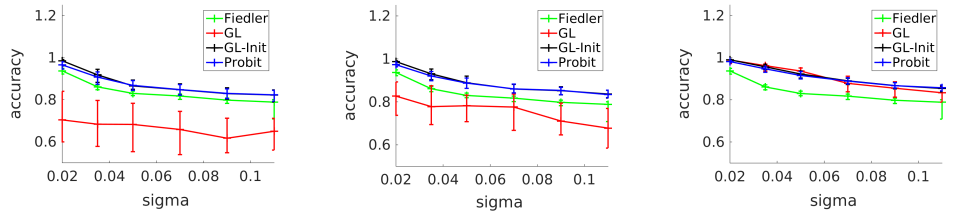
$$\begin{aligned} -x\int_{-\infty}^{x}\exp(\frac{-t^2}{2})dt - \exp(\frac{-x^2}{2}) &= -x\int_{-\infty}^{x}\exp(\frac{-t^2}{2})dt + \int_{-\infty}^{x} t\exp(\frac{-t^2}{2})dt \\ &= \int_{-\infty}^{x}(t-x)\exp(\frac{-t^2}{2})dt < 0 \end{aligned} \tag{3}$$

and the proof is complete. ∎

The probit MAP estimator thus has a considerable computational advantage over the Ginzburg-Landau MAP estimator, because the latter is not convex and, indeed, can have large numbers of minimizers. We now discuss numerical results designed to probe the consequences of convexity, or lack of it, for classification accuracy. The purpose of these experiments is not to match state-of-art results for classification, but rather to study properties of the MAP estimator when varying the feature noise and the percentage of labelled data.

We employ the two moons and the MNIST $(4, 9)$ data sets. The methods are evaluated on a range of values for the percentage of labelled data points, and also for a range of values of the feature variance σ in the two moons dataset. The experiments are conducted for 100 trials with different initializations (both two moons and MNIST $(4, 9)$) and different data realizations (for two moons only). In Figure 2, we plot the median classification accuracy with error bars from the 100 trials against the feature variance σ for the two moons dataset. As well as Ginzburg-Landau and probit classification, we also display results from spectral clustering based on thresholding the Feidler eigenvector. The percentage of fidelity points used is 0.5%, 1%, and 3% for each column. We do the same in Figure 3 for the 4 -9 MNIST data set against the same percentages of labelled points.

The non-convexity of the Ginzburg-Landau model can result in large variance in classification accuracy; the extent of this depends on the percentage of observed labels. The existence of sub-optimal local extrema causes the large variance. If initialized without information about the classification, Ginzburg-Landau can perform very badly in comparison with probit. On the other hand we find that the best performance of the Ginzburg-Landau model, when initialized at the probit minimizer, is typically slightly better than the probit model.



(a) Fidelity = 0.5% (b) Fidelity = 1% (c) Fidelity = 3%

FIG. 2. *Classification accuracy of different algorithms for Two Moons Dataset compared with σ and percentage of labelled nodes, with $N = 2,000$. The algorithms used are: Ginzburg-Landau MAP estimator with random initialization, Ginzburg-Landau with initialization given by probit model, probit MAP estimation, and spectral clustering (thresholding the Fiedler vector). For each trial, we generate a realization of the two moons dataset with given σ and select randomly a certain percentage of nodes as fidelity, and a total of 50 trials are run for each combination of parameters. We use spectral projection with number of eigenvectors $Neig = 150$. We plot the median accuracy along with error bars indicating the 25 and 75-th quantile of the classification accuracy of each method. We set $\gamma = 0.1$ for the probit model, and $\gamma = 1.0$, $\epsilon = 1.0$ for Ginzburg-Landau.*

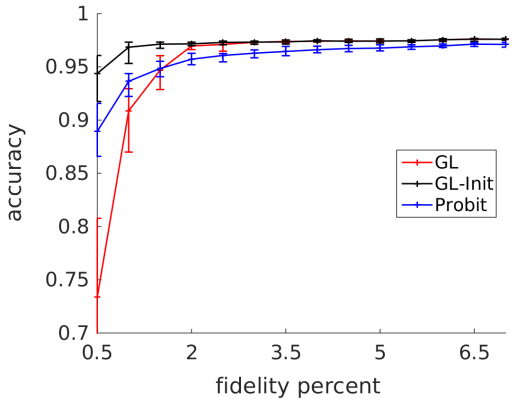


FIG. 3. *Classification accuracy of different algorithms for the 4-9 MNIST dataset versus percentage of labelled nodes. The algorithms used are: Ginzburg-Landau with random initialization, Ginzburg-Landau with initialization given by probit model, probit MAP estimation. For each trial, we select randomly a certain percentage of nodes as fidelity, and a total of 50 trials are run. We use spectral projection with number of eigenvectors $Neig = 300$. We plot the median accuracy along with error bars indicating the 25 and 75-th quantile of the classification accuracy of each method. We set $\gamma = 0.1$ for the probit model, and $\gamma = 1.0$, $\epsilon = 1.0$ for Ginzburg-Landau.*

We note that the probit model is convex and theoretically should have results independent of the initialization. However, we see there are still small variations in the classification result from different initializations. This is due to slow convergence of gradient methods caused by the flat-bottomed well of the probit log-likelihood. As mentioned above this can be understood by noting that, for small gamma, probit and level-set are closely related and that the level-set MAP estimator does not exist – minimizing sequences converge to zero, but the infimum is not attained at zero.

REFERENCES

[1] F. R. Chung, *Spectral graph theory*, vol. 92, American Mathematical Soc., 1997.
[2] U. Von Luxburg, M. Belkin, and O. Bousquet, *Consistency of spectral clustering*, The Annals of Statistics, (2008), pp. 555–586.